CSE8803-EPI

Project Proposal

Te-Kai Chen
Computational Science and Engineering
Georgia Institute of Technology
Atlanta GA US
tchen483@gatech.edu

Chen Chen
College of Computing
Georgia Institute of Technology
Atlanta GA US
chesterchen@gatech.edu

1 Formal problem definition

This project will concentrate on evaluating the machine learning based model and non-machine learning based model for the purpose of performance comparison. More specifically, we will choose three to five modes for both machine learning model and non-machine learning base model to analyze and visualize their advantages and disadvantages in different situations. It is significant to indicate the better model describing Covid-19 and what range of data would be considered as a top priority to improve performance and accuracy.

2 Models Description

2.1 Machine Learning based models

We are going to use four machine-learning-based models to solve the forecasting problem. We use the time-series model to deal with this problem since our objective is to estimate Covid-19 mortality and the average cases, which is a more likely time-related problem. The models are the standard Recurrent Neural Network (RNN) [1], Echo State Network (ESN) [2], Long Short-Term Memory (LSTM) [3], and Gated Recurrent Unit (GRU) [4]. RNN is a model which mainly has the input of sequential data. The difference between RNN and the typical feed-forward network is that RNN will consider the current information and the previous inputs. Hence, RNN can perform better than the regular feed-forward network with this characteristic when time is a crucial factor in data. The ESN model is a modification of RNN having a reservoir that randomly connects nodes inside the hidden layer. With this modification, ESN can solve some chaotic situations in the network. LSTM and GRU are also the same ideas as RNN but different from the inner blocks of the node. The LSTM block introduces the concept of gate preserving the portion of the information with current and previous information. Due to the ideas, it can help the model to learn whether the previous information is important or not. GRU is a modified version of LSTM because it preserves the gate's idea but replaces the gates with an update gate and reset gate. This modification can help the network with fewer parameters and make the training faster and more stable.

2.2 Non-Machine Learning based models

Among those time series models, there are six different models to deliver the significant results. We are going to select four models to address the relationship with Covid-19. AIRMA, auto-regressive Integrated Moving Average [5], Hot-Winters Additive Model (HWAAS)[6], TBAT[7], Prophet[8]: Automatic Forecasting Procedure [Ref]. In simple terms, AIRMA assume a linear correlation between times series values and attempt to estimate the dependencies. One of the benefits of ARIMA can be performed in an automated way to eliminate the prediction inaccuracy. HWAAS is extend from Holt's exponential smoothing, adds the seasonality factor to the forecast, which is best for data with trend that do not fluctuate or increase. TBAT is an estimation method to model non-integer seasonal frequencies, the method can be applied to wide range of times series problems. The advantages of TBAT framework that include a various of parameter with the better forecasts in time series data. Last but not least, Prophet is originally developed by Facebook to predict business-related outcomes. In order to forecast, Prophet uses an easily decomposable time-series model. Those models used to predict the Covid-19 cases. Besides, for training the validation purpose, in terms of the evaluation metric, RMSE (root mean squared error) would be the indicator to each of the approaches. Throughout the project, we will use methods list above to address the results and compare each of them with Covid-19 dataset (NYT Times).

3 Evaluation and Metrics

We are going to evaluate our model in the following situations. In temporal view, the first one is to use the whole dataset to predict our targets. The second one is to use the recent data to predict our targets. With the difference of scope, we can understand the ability of each model. In spatial view, we can respectively train our model with different data of states, prisons, and colleges. We can explore and find out the advantages of different models with these evaluations

Our metrics will be accuracy (median absolute error) and Root-Mean-Square-deviation (RMSE) to evaluate our models and minimize the error to find the best parameters. Furthermore, since each model may perform better in specific situations, we will also state the advantages and disadvantages of each model from different temporal views.

We will evaluate methods along with the NYT dataset. Since dataset contains daily information regarding several important indexes. For instance, the number of confirmed Convid-19 cases, the number of probable Covid-19 cases and the number of death due to Covid-19. Performance results of each model will be presented in forms of plots and graphs, the aim of the evaluation is to compare time series methods regarding predicting the percentage active cases of Covid-19. Furthermore, (RMSE) will be applied to predictive results and approaches. In addition, we will compare methods under different circumstances, which helps us to instantiate the model for certain situations.

4 Dataset Description

Dataset is recommended from CSE8803-EPI course piazza. The New York Times is releasing data with cumulative counts of coronavirus cases in the United States, this not only provides us with another form of visualization, but it also uses a formal, statistical model underneath.

We are going to choose NYT as our datasets, which contains the number of cases in various parts of the country. One of the advantages of NYT dataset is that the daily updated data can help us to evaluate the performance of the forecasting model. Datasets contains several files that are considered as EXCEL format such as us-counties-recent.csv, us-counties.csv, us-states.csv and us.csv. each excel file include various indexes, “date”, “county”, “state”, “fips”, “cases”, “deaths”, “confirmed_case”, “probable_case” and “probable_deaths”. The average of rows for each excel file or so is approximately 3000 rows.

5 Expectation

At the end of the semester, we expect to identify the benefit and drawback of each machine learning based models and non-machine learning models. For instance, using the root mean error (RMSE) to assess the performance of each time series model, compares the different approaches in different situations. Other than that, we will indicate performance results between models and point out which model could be adequate for Covid-19 prediction in different circumstances. With the result, we wish to find out the best range of data for long-term forecasting and short-term forecasting of our targets in both local (states and colleges) and global (US).

6 Work Division

The project would be completed by two members, Te-Kai Chen and Chen Chen. The main goal of our project is to analyze and compare the performance of machine learning models and non-machine learning models and figure out what models or methods could be suitable for analyzing Covid-19.

Te-Kai Chen: Responsible for using four machine learning based models, RNN, ESN, LSTM, GRU, to forecast the mortality and percentage of active cases per population.

Chen Chen: Responsible for using non-machine learning to compare the time series methods and forecast the mortality and percentage of active cases per population. More specially, utilizing the concept of AIRMA, HWAAS, TBAT, Prophet and Prophet to analyze NYT datasets.

Collaboration: In this project, among those four different statistical and machine learning-inspired time series methods, figure out which methods are well-developed for monitoring Covid-19 cases. Moreover, using visualization tools to generate graphs or line charts to understand the relationship between models and Covid-19, finalizing the models with their advantages and disadvantages.

Our expected timeline of activities:

Brainstorming and topics selection – Sep.28 ~ Sep.29

Drafting, proposal initialing – Oct.01 ~ Oct.04

Model training and analyzing – Oct.15 ~ Oct.30

Output visualization and prediction – Nov.5 ~ Nov.15

Summary and final project completion – Nov.16 ~ Nov.25

Presentation preparation – No.26 ~ Dec.1

REFERENCES

[1] Hewamalage H., Bergmeir C., Bandara K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions. Int. J. Forecast. 37, 388–427.

[2] Jaeger H, Haas H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science. 2004 Apr 2;304(5667):78-80.

[3] Yudistira N. COVID-19 growth prediction using multivariate long short term memory. arXiv. 2020;14(8):1–8.

[4] Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, and Bengio, Yoshua. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[5] Jamal Fattah, L. E. (n.d.). Forecasting of demand using ARIMA model - Jamal Fattah, Latifa Ezzine, Zineb Aman, Haj El Moussami, Abdeslam Lachhab, 2018.

[6] Yar, M., & Chatfield, C. (2002, April 23). Prediction intervals for the Holt-Winters forecasting procedure.

[7] Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015, June 09). Time series analysis for psychological research: Examining and forecasting change.

[8] Taylor, B. S., & Letham, B. (2021, March 01). Prophet: Forecasting at scale.